Use Reinforcement Learning to Choose Activities for Preparing to Quit Smoking

How Effective a Reinforcement Learning Model is for Choosing Activities that Optimizes the Likelihood that Users Return to the Next Session and the Effort Users Spend on Their Activities?

TU Delft Delft University of Technology

1. Background

- Societal: premature deaths cause by unhealthy behaviors such as smoking
- Psychological: difficult to quit smoking alone
- Technical: eHealth apps, reinforcement Learning

- conversational virtual coach
  - suggest quit-smoking-prep activities
- data: 5 sessions
  - before : usefulness beliefs about nine competencies (e.g. self-efficacy), energy level , busyness level, etc.
  - after: effort spent, dropout, etc.

2. Methodology

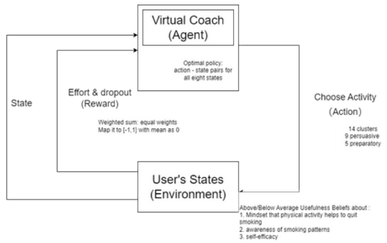


Figure 1: The reinforcement learning model with which the virtual coach chooses an activity cluster based on the users' current state and the optimal policy and then gives the reward and the next state of the user.

Formulate the reinforcement learning model:

- States:
  - from original data: 11 features with numeric values
  - need to reduce state space
  - turn each into binary by mean
  - select 3 by G algorithm
  - 8 states in total: e.g. [101]
- Actions:
  - 14 activity clusters
- Reward function:
  - effort spend & likelihood of return to next session
  - not really correlated: Cohen's kappa -0.04
  - weighted sum of both goals
  - map it to [-1, 1] with 0 as mean
- discount factor:
  - 0.85: not discourage users by low reward in next or near sessions
- optimal policy
  - optimal action for each state

3.1: How well can states derived from the features predict behavior ?

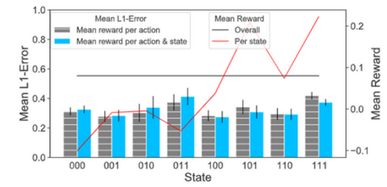


Figure 2: Left Axis: the mean prediction error per state based on 1) only the action and 2) the action and the state. Right Axis: the mean reward of each state and overall

Setup.

- compare the mean reward based on 1) only action with on 2) action and state
- leave-one-out cross validation
- L1-mean error with its 95% credible interval (CI)

Results.

- average reward: 0.08 for all states
- [111] : 0.22, and [101] : 0.21
- [000] : -0.1, and [011] : -0.05
- CIs overlap for most of the states
- state[111] gives better prediction

3.2: How well can the states predict states ?

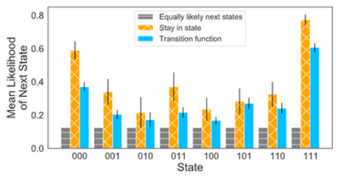


Figure 3: The mean likelihood of a transition to a next state by 1) assigning equal probability to all next states, 2) assuming stay in the current state, and 3) using probability based on estimated transition function of RL model

setup.

- Compare three approaches of predicting the next states by their mean likelihood and 95% CIs
- Leave-one-out cross validation

Results:

- the distribution of next states is not uniform.
- [000], [011] and [111]
  - no CI overlaps,
  - high mean likelihood of staying current
  - tend to stay at current in states with very high or very low mean reward per state

3.3: What is the effect of performing multiple optimal actions on states?

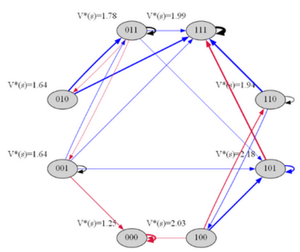


Figure 4: Transition probabilities under π^*. Only transition with a probability of at least $\frac{1}{|S|}$ are shown. Red arrows means transitions to next states with a lower value. Black arrows means transitions to states with the same value. Blue arrows means transitions to states with higher value. The thicker the line, the higher the probability of the transition.

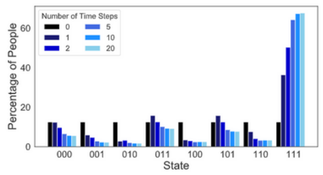


Figure 5: Percentage of people in each state after following π^* for some certain number of steps.

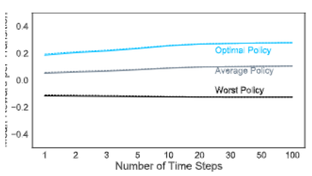
Setup.

- obtain optimal policy by value iteration
- simulate users with even distribution of all the states follow this policy for some certain number of time steps

Results:

- state [101] has the highest value V*(s), 2.18
- users tend to move to states with higher value, such as [011], [111], [110] and [101]
- users tend to stay in states with higher value, such as 44% for [101] and 86% for [111]
- red arrows suggest chances of moving from higher value states to lower value states
  - users in [000] tend to stay with probability of 55%
  - users in [100] have tendency to go to [000]
- users tend to move to state [111], highest average reward per state high value
- still 5.6% of users stay in state [000]

3.4: Compare optimal policy with non-optimal for the effect on behavior



igure 6: Mean reward per transition over time for three policies. he initial populations are the state distribution of all people(solid ne) or of only the people with a reward in the lowest 25%(dashed ne) for the first session

Setup.

- average policy: each action was taken an equal amount of times at all time steps
- worst policy: the least optimal, lowest reward

Results:

- the optimal policy gives the best mean reward per transition
- increased from 0.20 to 0.29 for the general population
- increased from 0.21 to 0.29 for the lowest 25%
- the average policy increased it by 0.05, compared to that of the optimal policy, 0.08

3.5: Compare the effect on behavior of optimal policy obtained by different reward functions

Weight	[000]	[001]	[010]	[011]	[100]	[111]	[110]	[111]
0.25	9	12	2	5	13	4	13	7
0.5	9	12	1	5	13	4	13	7
0.75	5	12	3	5	13	4	7	6

Table 1: Optimal policies obtained from different weighted sum of the effort and the return likelihood by effort being 1) 25%, 2) 50% and 3) 75% of the sum.

Setup:

- reward is calculated with the same approach except the weight distribution is different: effort being 1) 25%, 2) 50% (original), and 3) 75% of the weighted sum
- state space does not change
- obtain optimal policies for the new reward functions
- transition graphs with values for each state
- percentage of people in each state after following the corresponding optimal policy.

Results:

- Different reward functions gives different optimal actions
- Relatively high similarity: [001], [011], [100] and [111] identical for all three approaches
- only state [010] and [111] optimal action is preparatory, others are persuasive
- transition graphs similar to original one
- move users to [111]
- original has lowest percentage for [000]

4. Conclusions and Limitations

Discussions:

- Better prediction for the behavior(reward), especially for [111]
- Users tend to stay in better states and also in the worst state.
- most people tend to move to better states, but still some left in the less good states.
- the optimal policy we obtained does result in better reward.
- changing weight distribution of the two goals result in different but similar optimal policies
- Reward relies less on weight distribution

Limitations:

- assumption of the relation of the two goals: weighted sum
- assumption that the transition function and reward function does not change
- potential inconsistencies and possible reasons
  - the second belief has negative impact on our goals
  - we did not test on real subjects, instead we did simulations

Author: Meng Zhang

Supervisor: Nele Albers

Responsible Professor: Willem-Paul Brinkman